Amazon Biobank - A community-based genetic database

Leonardo T. Kimura[1], **Ewerton R. Andrade**[1,2], **Tereza C. Carvalho**[1], **Marcos A. Simplício Jr.**[1]

[1]Escola Politécnica - University of São Paulo (USP)

[2]Universidade Federal de Rondônia (UNIR)

lkimura@larc.usp.br, ewerton.andrade@unir.br

terezacarvalho@usp.br, mjunior@larc.usp.br

Abstract. *In regions like the Amazon Rainforest, there is much unexplored biodiversity data that could potentially be used to promote innovative biotechnology developments. Building a biobank with such genetic data is, however, a challenge. One reason is that existing repositories (e.g., NCBI) lack clear incentives for collaboration. Aiming to tackle this issue, and promote a biodiversity-based economy in the Amazon region, in this work we present a prototype for the Amazon Biobank, a community-based genetic database. Leveraging blockchain, smart contracts, and peer-to-peer technologies, we build a collaborative and highly scalable repository. It also enables monetary incentives for users who insert, store, process, validate and share DNA data.*

1. Introduction

Biotechnology is defined as transforming natural products, like living organisms or substances, into commercial products, such as medicine or cosmetics [Pearce and Moran 1994]. The number of biotechnology products has been growing in the last year, generating sustainable inventions such as biodegradable plastic [Lokko et al. 2018]. More generally, this trend has the potential to generate considerable benefits in fields such as agriculture, pharmaceuticals, and cosmetics, improving production efficiency while limiting environmental impacts [Soetaert and Vandamme 2006].

There are today many repositories that provide genetic data required for biotechnology research. Examples include NCBI (National Genetic Heritage Management System), in the United States, and SisGen (National System for Managing Genetic Assets and Associated Traditional Knowledge), in Brazil. Also, the UNDP (United Nations Development Programme) has recently described a global initiative to enable traceability and benefit-sharing of genetic data through blockchain technology [UNDP 2021].

However, although the volume of DNA data inserted in genome-oriented repositories increases annually [NCBI 2021], most of the species in remote areas like the Amazon Rainforest are not yet cataloged [Cardoso et al. 2017]. One important reason is the difficulty to collect DNA from the region, specially due to the high costs of exploratory expeditions to the area. Another cause is the lack of clear incentives for inserting genomic data into such systems, as in most cases there is no monetary remuneration for this effort. Therefore, even people with easy access to genomic data, like local residents, would not feel compelled to contribute to the system. Instead, they are more likely to engage in

more lucrative activities, like wood and gold extraction. Unfortunately, the negative environmental impacts and loss of biodiversity resulting from such activities are prone to undermine any opportunity of profiting from a biotechnology-based economy. Reverting this scenario is, thus, of utmost importance.

Aiming to tackle these issues, we hereby describe a prototype for the Amazon Biobank initiative [Kimura et al. 2021], a community-based genetic database. Built upon Blockchain and peer-to-peer (P2P) technologies, it implements monetary incentives to insert, store and validate DNA data. Users can collaborate with the system not only by providing DNA data, but also computational power and bandwidth, creating a highly scalable environment. In addition, Amazon Biobank allows the adequate sharing of the produced wealth among participants through smart contracts registered in a Blockchain for easy traceability and auditability.

2. Background

This section describes the two main building blocks of the Amazon Biobank system: blockchain, and peer-to-peer file-sharing networks.

Blockchain is frequently used to enable the exchange of digital assets among parties who do not trust each other. One classical example is Bitcoin [Nakamoto 2008], a system based on digital coins that are exchanged among different untrusted parties. Bitcoin uses a blockchain as a distributed time stamp authority, creating a reliable log of transactions while dealing with attempts of double-spending. [Swan 2015] describes the technology as a public ledger of all Bitcoin transactions executed throughout their existence. This ledger is constantly growing through miners, who add new blocks to record the most recent transactions.

In the Amazon Biobank, we use blockchain for enabling two types of digital assets. The first are tokens referring to biodiversity knowledge, represented by DNA Sequences and their corresponding metadata. Specifically, the goal, in this case, is to enable data owners to define smart contracts that control the access to their data, as well as how different entities participating in the collection, processing, and distribution of this data are remunerated when this data is accessed. The second are virtual coins (named “biocoins”), which are minted in response to the insertion of assets into the system and used for purchasing services (e.g., the right to access some piece of data). Such biocoins may also be traded freely, similarly to other digital currencies in literature.

Some blockchains in use today are called ”non-permissioned”, meaning that it is fully accessible by any interested user, so the system is monitored by anyone and do not belong to anyone [Swan 2015]. One example is the blockchain employed in Bitcoin itself. Another category of blockchain is permissioned blockchains, in which only some authorities control the network operations and access. In this case, it is possible to assign a role to each member of the network, define who can write or access information, and admit and expel members from the system. This ends up facilitating many internal procedures of the blockchain’s operation; for example, it enables the deployment of simpler and more efficient consensus mechanisms, while misbehavior can involve real-world sanctions besides on-chain treatment. Since this kind of control and high efficiency are required in the target system, we adopt a permissioned blockchain architecture based on the Hyperledger Fabric framework. [Androulaki et al. 2018]

Finally, another cornerstone component of the system is a peer-to-peer file system, i.e., a method of distributing files among users (also called peers) without the need of keeping files on a specific, high availability server [Cohen 2003]. This is relevant in our target scenario because of the typically very large size of raw DNA data, which commonly reaches the order of hundreds of gigabytes. The specific solution adopted in the Amazon Biobank prototype hereby described is the BitTorrent protocol [Cohen 2003]. In a nutshell, BitTorrent allows large data pieces to be split into several chunks of arbitrary size, and then downloaded from different peers; it also provides strong integrity guarantees, by including the hash of each chunk in the metafile that enables the download of such chunks (called a torrent file). Both properties lead to a highly scalable and reliable system when dealing with large amounts of data and many users.

2.1. Related Works

Several studies explore the use of blockchain for registering and distributing DNA data. For instance, [Ozercan et al. 2018] and [Thiebes et al. 2020] discuss the potential that blockchain holds for genomics, whereas companies like Nebula Genomics [Grishin et al. 2018] and Encrypgen [Encrypgen 2021] use the technology in their business model. However, these companies typically act as gatekeepers, centralizing the system and removing many of the advantages of decentralization (e.g., scalability and availability). In addition, they prioritize mainly human genetics, giving little emphasis on biodiversity as an asset.

Currently, the UNDP is conducting a project based on Blockchain to improve genome resource traceability and benefit-sharing [UNDP 2021]. Their major goal is to implement the Nagoya Protocol, adopted in 2010 [Buck and Hamilton 2011]. Called Global ABS Tracker, the project is presently in the prototype phase, and the UNDP plans to deploy an alpha version by 2022. The system focuses on natural products, like plants, and not on genetic data per se; besides, unlike our work, it does not handle the (collaborative) storage of genomic data. The project also needs global coordination between countries to support its operation, something that is still in progress.

3. System Overview

As shown in Fig. 1a, the system comprises several players, namely: (1) collectors, responsible for collecting raw DNA data and inserting it into the system; (2) distributors, responsible for storing and sharing DNA data; (3) processors, responsible for transforming the raw DNA data into a more useful, annotated DNA form; (4) buyers, that pay biocoins to purchase the right to access DNA data pieces aiming to create biotechnological products; and (5) the federation, responsible for maintaining the Biobank system. In what follows, we describe the main operations performed by those players.

3.1. Create and upload a DNA sequence

One of the system's primary operations is the registration of raw DNA data (Fig. 1a, operation 1). A Collector (e.g., a resident of the Amazon region) gathers the data using low-cost PCR equipment, extracting a raw instrument signal or a DNA/RNA sequence read. The data is then encrypted to preserve the confidentiality of its contents during distribution. The decryption key can can be stored by the federation, providing backup and usability improvements, or by collector, keeping the DNA data even more confidential.

As the data owner, the collector inserts it into the system, defining the conditions for its utilization, like which rules (e.g., royalties) apply, and how other players in the chain (e.g., Processors) are rewarded when the content is purchased. Those rules are enforced via smart contracts registered in the blockchain.

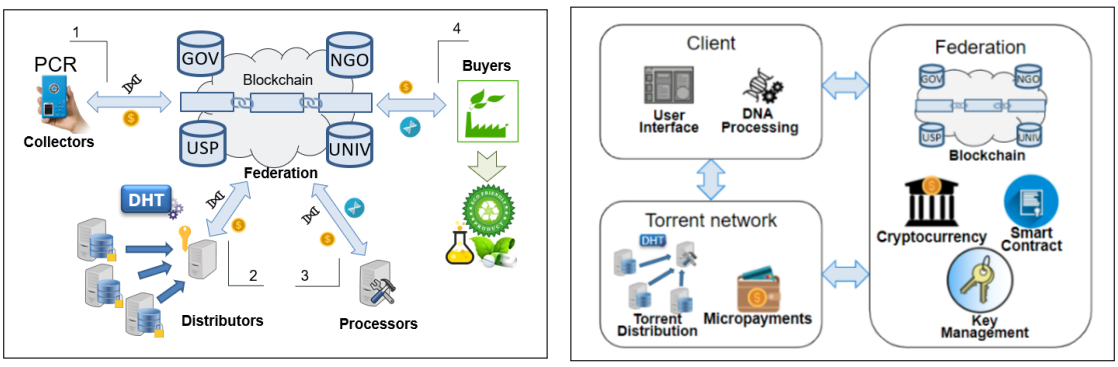
The insertion of DNA data into the system may also involve Curators (typically, biologists), who may analyze each inserted DNA and decide if the data is novel or redundant considering existing data. Collectors who insert new information into the system are expected to be rewarded with freshly minted biocoins, as a reward (and incentive) for growing the Biobank's data relevance.

3.2. Distribute a DNA sequence

Since DNA sequences are usually large, they should not be directly stored in a blockchain. Instead, the DNA is divided into several chunks as per the BitTorrent protocol, and only their magnet links [Xinxing et al. 2016] are stored in the Blockchain (Fig. 1a, Operation 2). Registered Distributors can then declare their interest in a given DNA data can then download, store and share the corresponding encrypted data chunks via BitTorrent, without gaining access to the plaintext content. The motivation for doing so is that Processors and Buyers who purchase the decryption key (see Sec. 3.3 and 3.4) are expected to also reward Distributors with biocoins, using a micropayment mechanism [Micali and Rivest 2002] for acquiring each data chunk.

3.3. Processing the DNA sequence

Raw DNA sequences are expected to be processed by interested parties, thus creating assembled DNA Sequences that can be uploaded back to the Biobank system (Fig. 1a, Operation 3). This task is performed by any party registered in the role of Processor, who must purchase biocoins for rewarding the Distributors who share the encrypted DNA data and for buying the corresponding decryption key. This procedure may require a considerable amount of processing power. Therefore, like Colletors, Processors who upload data into the system should be rewarded accordingly when the corresponding data is purchased by Buyers. The access to this processed data is also governed by smart contracts. Also, as an incentive and to ensuring a healthy presence on Processors in the system, Processors may receive newly minted biocoins when they register, and also have their download fees refunded after registering assembled sequences from raw DNA.



(a) Main players and operations

(b) Basic architecture

Figure 1. Amazon Biobank overview.

3.4. Purchasing the DNA data

Both raw and processed/validated DNA sequences are available for download by interested parties, e.g., from the industry or academia (Fig. 1a, Operation 4). Once again, such downloads are expected to be paid in biocoins, so that entities that took part in the generation and distribution of the data (Collectors, Processors, and Distributors), can be remunerated according to the corresponding smart contract. This purchase is registered into the Blockchain, so this entry serves as proof that the buyer agrees with the terms of use for the purchased piece of data, and then receive the corresponding data decryption keys. Even when royalties are involved, it is expected that such costs would be much lower than an Amazon expedition to collect the DNA data of interest.

We note that this access control mechanism, via data encryption, does not prevent decryption keys from being exposed outside the Biobank system. There should be little incentive for doing so, though: after purchasing the decryption key, exposing it only increases availability for competitors rather than bringing any benefit. Also, typical Buyers should be interested in having the purchase of DNA registered in the Blockchain, since this provides data traceability (important for research in the biology field) and enables a "helping the Amazon Forest" seal to be included in products created using the Biobank's data. Nevertheless, in case of intellectual property infringement, the Biobank's underlying blockchain is still useful for resolving disputes: it gives an approximate time frame of when the DNA appeared in the Biobank, which can then be compared with the release of products that are claimed to be derived from such data.

4. Architecture

Fig. 1b shows the main components of the system: a web-based application layer providing a graphical user interface for interacting with the system; a BitTorrent-based data sharing sub-system for exchanging DNA data; and the back-end Federation, built using Hyperledger for recording biocoins, magnet links, and smart contracts.

The data sharing layer is maintained by distributed machines participating in BitTorrent swarms, one for each DNA data registered in the system. The corresponding magnet links can be generated by any BitTorrent client, given the target (encrypted) DNA data. A dedicated client application based on qBittorrent (`https://www.qbittorrent.org`) is currently under construction for facilitating this task.

The web-based application layer handles the user access to the Biobank. Through it, users can interact with the Blockchain and BitTorrent networks, make queries and write data into the system. Communication is done via a secure tunnel, so the magnet links and the corresponding decryption keys can be registered safely. At that time, users can pick a pre-built smart contract to be associated with their data; they can also adjust some of its contract's parameters, such as the selling price.

Finally, the blockchain layer is maintained by a Federation of universities and NGOs interested in running the Biobank. These organizations support the cryptocurrency modules, smart contracts, and access control (including registration of new users and management of decryption keys). Federation members are assumed to be trustworthy, although extra mechanisms may be employed to avoid misbehavior if desired (e.g., keys could be handled via secret sharing schemes [Beimel 2011] to avoid exposure by any single entity).

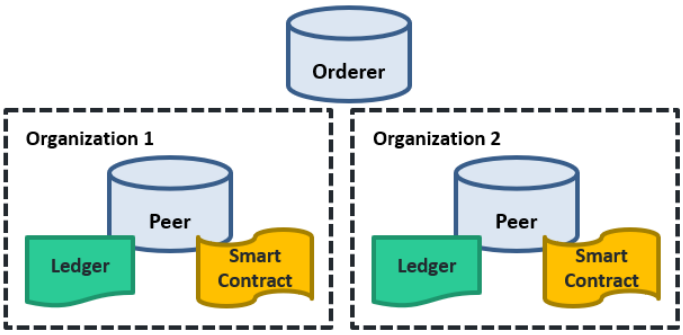


Figure 2. Blockchain test network infrastructure.

5. Prototype and proof-of-concept

In our prototype, the Hyperledger Fabric is used to build a Federation with one orderer and two peers, each of which having a storage ledger and chaincode (or smart contract). Fig. 2 depicts the resulting network. These ledgers and chaincode are consistent with each other, so that the Biobank system can run in unity, and the organizations can monitor each other's activities. To visualize blockchain data, we configured a Hyperledger Explorer application [Dhillon et al. 2017]. That way, users can view all transactions sent to the system and all blocks approved. Any interested auditor can also monitor the blockchain to verify its append-only property, vouching for its integrity in case of disputes. For example, monitors could inform the age of a given DNA data during a dispute involving intellectual property infringement, or detect attempts (by federation members) of changing the ownership of a given magnet link after registration.

The web-based application layer was built using express.js. Interactions with the system are authenticated using x509 certificates, which gives access to the blockchain layer using the Hyperledger Fabric SDK.

6. Demonstration

The prototype supports the essential operations of the Biobank system: inserting DNA data via magnet links, searching them by metadata, purchasing biocoins and decryption keys, running smart contract functionalities, among others. Fig. 3 illustrates how Hyperledger Fabric can be used to see the system contents (namely, a piece of raw DNA data registered by a Collector).

The prototype implementation is available at https://github.com/amazon-biobank/biobank. It comprises all the documentation about the installation, basic requirements, and some descriptions of API. We deployed the prototype in two places: an on-premise server, and a local machine, used for testing purposes. Our demonstration at SBSeg needs both of these computers, but there is no specific requirement for these machines. In addition, connection to the Internet is necessary.

The demonstration includes registering magnet link for DNA data in the system, configuring smart contract parameters, searching for metadata, and purchasing DNA data. Hyperledger Fabric registers all the transactions in the blockchain, so, using the Hyperledger Explorer application, we can consult these transactions for further analyses. A preliminary video of the demonstration is available online at https://bit.ly/3AyTBlU.

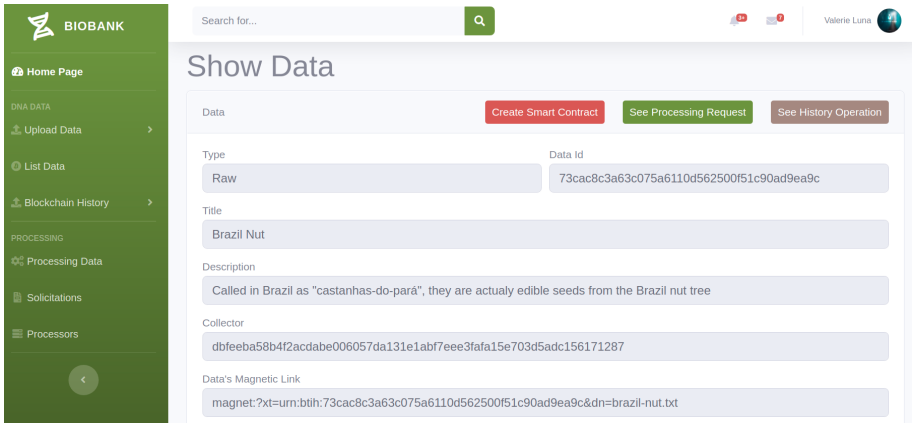


Figure 3. Inserting a DNA register in the system.

7. Conclusion and Future Works

There are several genetic databases worldwide, however, although the volume of DNA data inserted in genome-oriented repositories increases annually, most of the species in remote areas like the Amazon Rainforest are not yet cataloged. Experts point out reasons such as difficulty to collect DNA from the region and the lack of clear incentives for inserting genomic data into these systems as the main reason for this event. Nevertheless, in areas such as the Amazon rainforest, several local communities with easy access to these data could contribute to the system as long as they are properly encouraged to do so instead of predatory activities like wood and gold extraction.

Therefore, we present a prototype for the Amazon Biobank initiative, a community-based genetic database. Built upon Blockchain and peer-to-peer (P2P) technologies, it implements monetary incentives to insert, store and validate DNA data. In this way, the local community may be paid to collaborate with the system providing DNA data and maintain forest biodiversity. This approach also enables other users around the world to store, process, and validate this data by sharing computational power and bandwidth, creating a highly scalable environment. This collaborative approach allows the University, Industries, and other institutions to use (even commercially) such data without the other entities in the process being financially harmed.

To enable monetary incentives and discourage misconduct, the proposed solution uses smart contracts registered in a Blockchain for easy traceability and auditability. At this point, our prototype implements some of Amazon Biobank major operations, validating the initiative's viability.

Acknowledgments: This work was supported by Ripple's University Blockchain Research Initiative, and by CNPq (grant 304643/2020-3).

References

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., et al. (2018). Hyperledger Fabric: a distributed operating system for permissioned blockchains. In *13th EuroSys*.

Beimel, A. (2011). Secret-sharing schemes: A survey. In *Coding and Cryptology*, pages 11–46, Berlin, Heidelberg. Springer.

Buck, M. and Hamilton, C. (2011). The nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. *Review of ECIEL*, 20(1):47–61.

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A. M., Bittrich, V., Celis, M., Daly, D. C., Fiaschi, P., Funk, V. A., Giacomin, L. L., et al. (2017). Amazon plant diversity revealed by a taxonomically verified species list. *Proc. of NAS*, 114(40):10695–10700.

Cohen, B. (2003). Incentives build robustness in bittorrent. In *Workshop on Economics of Peer-to-Peer systems*, volume 6, pages 68–72. Berkeley, CA, USA.

Dhillon, V., Metcalf, D., and Hooper, M. (2017). The Hyperledger project. In *Blockchain enabled applications*, pages 139–149. Springer.

Encrypgen (2021). `https://encrypgen.com/`. Accessed in 29-06-2021.

Grishin, D., Obbad, K., Estep, P., Quinn, K., Zaranek, S., Zaranek, A., Vandewege, W., Clegg, T., César, N., Cifric, M., and Church, G. (2018). Accelerating genomic data generation and facilitating genomic data access using decentralization, privacy-preserving technologies and equitable compensation. *Blockchain in Healthcare Today*.

Kimura, L., Andrade, E., Carvalho, T., and Simplicio, M. (2021). Amazon Biobank: sustainable development built upon rainforest's biodiversity. In *Planetary Health Annual Meeting and Festival*. `https://bit.ly/3eyLht9`.

Lokko, Y., Heijde, M., Schebesta, K., Scholtès, P., Van Montagu, M., and Giacca, M. (2018). Biotechnology and the bioeconomy—towards inclusive and sustainable industrial development. *New biotechnology*, 40:5–10.

Micali, S. and Rivest, R. L. (2002). Micropayments revisited. In *Cryptographers' Track at the RSA Conference*, pages 149–163. Springer.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260.

NCBI (2021). GenBank and WGS Statistics. `https://www.ncbi.nlm.nih.gov/genbank/statistics`. Accessed on 29-06-2021.

Ozercan, H. I., Ileri, A. M., Ayday, E., and Alkan, C. (2018). Realizing the potential of blockchain technologies in genomics. *Genome Research*, 28(9):1255–1263.

Pearce, D. W. and Moran, D. (1994). *The economic value of biodiversity*. Earthscan.

Soetaert, W. and Vandamme, E. (2006). The impact of industrial biotechnology. *Biotechnology Journal: Healthcare Nutrition Technology*, 1(7-8):756–769.

Swan, M. (2015). *Blockchain: Blueprint for a new economy*. O'Reilly Media, Inc.

Thiebes, S., Kannengießer, N., Schmidt-Kraepelin, M., and Sunyaev, A. (2020). Beyond Data Markets: Opportunities and Challenges for Distributed Ledger Technology in Genomics. *Proc. of the 53rd Hawaii Int. Conf. on System Sciences*, 3:3275–3284.

UNDP (2021). A pilot to improve genetic resources traceability through blockchain technology launched by the UNDP GEF Global ABS project. `https://bit.ly/3hnMqEh`. Acessed on 29-06-2021.

Xinxing, Z., Zhihong, T., and Luchen, Z. (2016). A measurement study on mainline DHT and magnet link. In *IEEE 1st Int. Conf. on Data Science in Cyberspace*, pages 11–19.